*Bike Renting*

Ankush Saha

# 

# Table of Contents

[1 Introduction 3](#_Toc16200197)

[1.1 Problem Statement 3](#_Toc16200198)

[1.2 Data 3](#_Toc16200199)

[2 Methodology 4](#_Toc16200200)

[2.1 Preparing Data 4](#_Toc16200201)

[2.2 Data Pre-Processing 4](#_Toc16200202)

[2.2.1 Missing Value Analysis 4](#_Toc16200203)

[2.2.2 Outlier Analysis 4](#_Toc16200204)

[2.2.3 Feature Selection 5](#_Toc16200205)

[2.2.4 Feature Scaling 5](#_Toc16200206)

[2.2.5 Sampling 6](#_Toc16200207)

[3 Model development 6](#_Toc16200208)

[3.1 Linear Regression 6](#_Toc16200209)

[3.2 Decision Tree 8](#_Toc16200210)

[3.3 Random Forest 8](#_Toc16200211)

[3.4 KNN Model 8](#_Toc16200212)

[4 Model Evaluation 9](#_Toc16200213)

[4.1 MAE 9](#_Toc16200214)

[4.1 MAPE 9](#_Toc16200215)

[5 Model Selection 10](#_Toc16200216)

[6 Output 10](#_Toc16200217)

[7 Visualization 11](#_Toc16200218)

[7.1 count of rent weekday wise 11](#_Toc16200219)

[7.2 count of rents holiday wise 12](#_Toc16200220)

[7.3 count of rents month wise 13](#_Toc16200221)

# 1 Introduction

## 1.1 Problem Statement

We are a bike rental company. We have collected the Historical data from your pilot project and now have a requirement to apply analytics for total number of bike rented in a day.

## 1.2 Data

Our task is to build a regression model which will predict the bike rental count based on given data. Below is a sample of the data set that we are using to predict number of bike rentals:

**day:**



# 2 Methodology

Any predictive modeling requires that we look at the data before we start modeling. First we prepare the data-set and we prepare the Data-set to feed into our model. In data preparation we perform actions like missing value analysis, outlier analysis, feature scaling, feature sampling.

## 2.1 Preparing Data

Before any kind of analysis we prepare the data. In this case we are removing dteday column as it is then split into yr, mnth, and weekday.

## 2.2 Data Pre-Processing

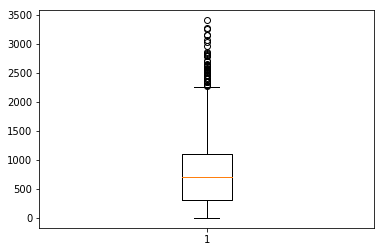
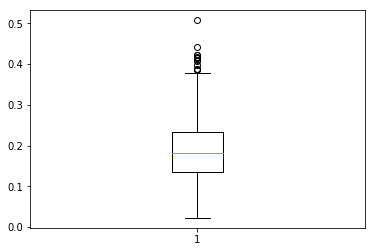
### **2.2.1 Missing Value Analysis**

Here we check if there is any missing value cell in the data set or not. If there is any missing value cell in the data set it may affect the model, and our model may not be efficient enough to predict correct value. So we need to either empty or impute those empty cells. If the number of empty cells in a column is more than 30 percent we can drop that variable as it may not add much value in our model. And if it is less than 30 percent we can impute them with either of any basic statistical process (i.e. mean, median, mode) or we can use KNN imputation method which one is best fit.

After our analysis we found out that in our data set there are no missing values present.

### **2.2.2 Outlier Analysis**

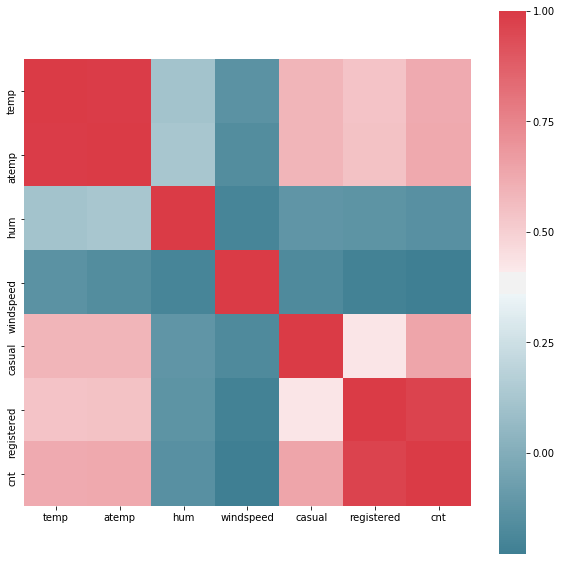
Sometimes in a column of data set there are some values which do not comply with general behavior of other data. These data are called outliers. These values may manipulate the behavior of the data set. And our model may not work accurately. To tackle this kind of situation we can either delete the row or impute the values with mean, median, mode or KNN imputation.



Boxplot 1: Windspeed

Boxplot : Casual

In our dataset we used boxplot analysis and find out that there are outliers were present in ‘windspeed’ and ‘casual’ column. And the number of outliers were 79 which is 10% of our data set so we have removed the outliers.



### **2.2.3 Feature Selection**

There might be some variables who are highly dependent to each other. So keeping both of them in our data set may create some partiality towards some features. So we can remove one of the variables.

In our dataset there were 2 types of variables. Continuous and categorical variables.

For continuous variables we have used collinearity check and plotted them, and we found out that temp and atemp are highly correlated to each other. And registered and count are highly correlated to each other. So we removed temp and registered.

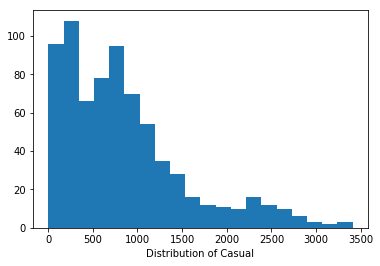
Figure : Correlation of continuous variables

For categorical variables we have used chi square test for every 2 variables. And found out that 'season','weekday','workingday','weathersit' are mostly correlated to other categorical variables. So we also removed them

### 2.2.4 Feature Scaling

Now every variable has their own scale of measurement. As we are considering only values and not the scales the values of the different variable may differ in very high range. It may cause to partiality towards higher valued variable. So we need to scale them up or town to a common scale.

Here in our data set we have variable named ‘casual’ which has higher value than other. So our model may get biased to this variable. So we need to scale this variable To achieve this we have plotted the variable and checked it’s distribution.



From this diagram we can say that the values are not normally distributed. So we have scaled it using standardization. To achieve this we have used below formula

### **2.2.5 Sampling**

First we need to divide our data set into two parts. First part on which we will develop our model and second part on which we will test our model how accurate it will predict.

So for sampling we have used random sampling technique which will randomly separate data into train and test data set for given size and for our data set we have taken 70% for train data and 30% for test data

After sampling train data size is (460 X 8)

And test data size is (192 X 8)

# 3 Model development

For prediction we need to develop a model by which we will predict the cab fare between two points. There are different types of model building technique available. For different kind of data set different techniques work accurately. So we will develop three models using three different techniques and will check which model is working more accurately, and we will feed our test data set to the best model and develop our prediction.

The three techniques we are goanna use are

1. Linear regression
2. Decision tree
3. Random Forest
4. KNN model

## 3.1 Linear Regression

In statistics, the linear regression model is used to predict some value using values of other dependent variables. Behind the algorithm it calculates the relation between the dependent variables with independent variables and calculate co-efficient for those dependent variables. Then while prediction it put the values of the dependent variables in the formula and calculate the dependent variable

First it makes a formula depending on the independent variables to calculate the dependent variable. The formula is as shown below

Where y = predicted value

= coefficient of the nth independent variable.

It also calculate std. Errors, t value and p value.

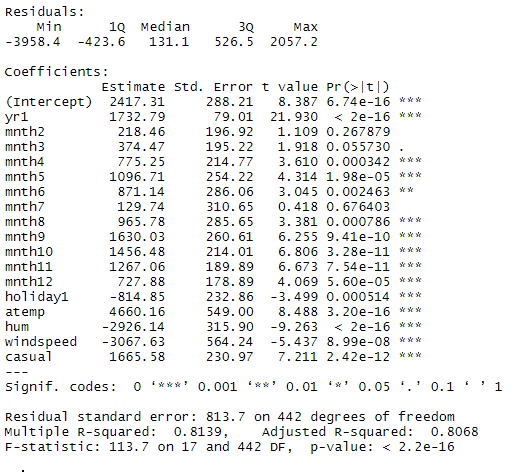
Std error provides estimated error associated with the estimations.

t-value is (estimation)/(std. Error).

The absolute value of t-value must be less than std error.

The p-value is a probability value. The p-value will be always between 0 and 1.

Lower p-value indicates statistically significant effect of predictor on dependent variable.



As we can see adjusted R-square value is .8068, means we can explain about 81% of the data using our multiple linear regression model. Which is very impressive.

## 3.2 Decision Tree

In [computer science](https://en.wikipedia.org/wiki/Computer_science), Decision tree learning uses a [decision tree](https://en.wikipedia.org/wiki/Decision_tree) (as a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling)) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in [statistics](https://en.wikipedia.org/wiki/Statistics), [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](https://en.wikipedia.org/wiki/Real_numbers)) are called regression trees.

## 3.3 Random Forest

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

## 3.4 KNN Model

KNN stands for K-Nearest Neighbor. It is a simple algorithm which store all available cases and classifies new cases based on a similarity measure. It is a very slow learning method. We don’t need to train our data set first.

We would give a input of number of neighbors we want to use for our model. it will pick the neighbors by measuring the distance from the data using below formula



# 4 Model Evaluation

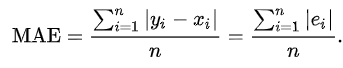
For Model Evaluation We use various techniques. We are building a regression model. For regression model we use methods like MAPE, MAE, RMSE.

In our dataset we have used MAPE and MAE.

## 4.1 MAE

Mean Absolute Error (MAE) is a measure of difference between two continuous variables. Assume X and Y are variables of paired observations that express the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. Consider a scatter plot of n points, where point i has coordinates (xi, yi)... Mean Absolute Error (MAE) is the average vertical distance between each point and the [identity line](https://en.wikipedia.org/wiki/Identity_line). MAE is also the average horizontal distance between each point and the identity line.

The Mean Absolute Error is given by:

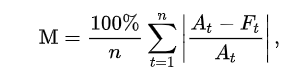
{\displaystyle \mathrm {MAE} ={\frac {\sum \_{i=1}^{n}\left|y\_{i}-x\_{i}\right|}{n}}={\frac {\sum \_{i=1}^{n}\left|e\_{i}\right|}{n}}.}

In Python MAE for linear regression =598.952, for decision tree = 666.66, for Random forest = 522.61 and for KNN = 9.968

In R MAE for linear regression = 668.7299973770903, for decision tree = 676.27, for Random forest = 443.70 and for KNN = 535.507

## 4.1 MAPE

Mean Absolute Percentage Error (MAPE)  also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in [statistics](https://en.wikipedia.org/wiki/Statistics), for example in [trend estimation](https://en.wikipedia.org/wiki/Trend_estimation), also used as a [loss function](https://en.wikipedia.org/wiki/Loss_function) for regression problems in [machine learning](https://en.wikipedia.org/wiki/Machine_learning). It usually expresses accuracy as a percentage, and is defined by the formula:



In R MAPE for linear regression = 19.46, for decision tree = 22.53, for Random forest = 18.87, and for KNN = 0.33

In Python MAPE for linear regression = 20.72, for decision tree = 21.9, and for Random forest = 14.06, and for KNN = 19.09

# 5 Model Selection

From all of the test we can clearly say that KNN is the best fit for the model in R and Random forest is best fit for the model in python. As its MAPE is very low from the other models.

So we will use KNN model for predicting cab fare of our test data when we will use R and we will use Random Forest when we will use python.

# 6 Output

For example output we have created a scenario where we have assumed the helpful parameters. Like year = 2011, month = December, it was not a holiday, normalized feeling temperature of the day = 0.474, normalized humidity was 0.627, normalized wind speed = 0.19, and casual users were 713.

Now we have fit these data in appropriate model in R and python i.e. KNN for R and Random Forest for python and extracted the output.

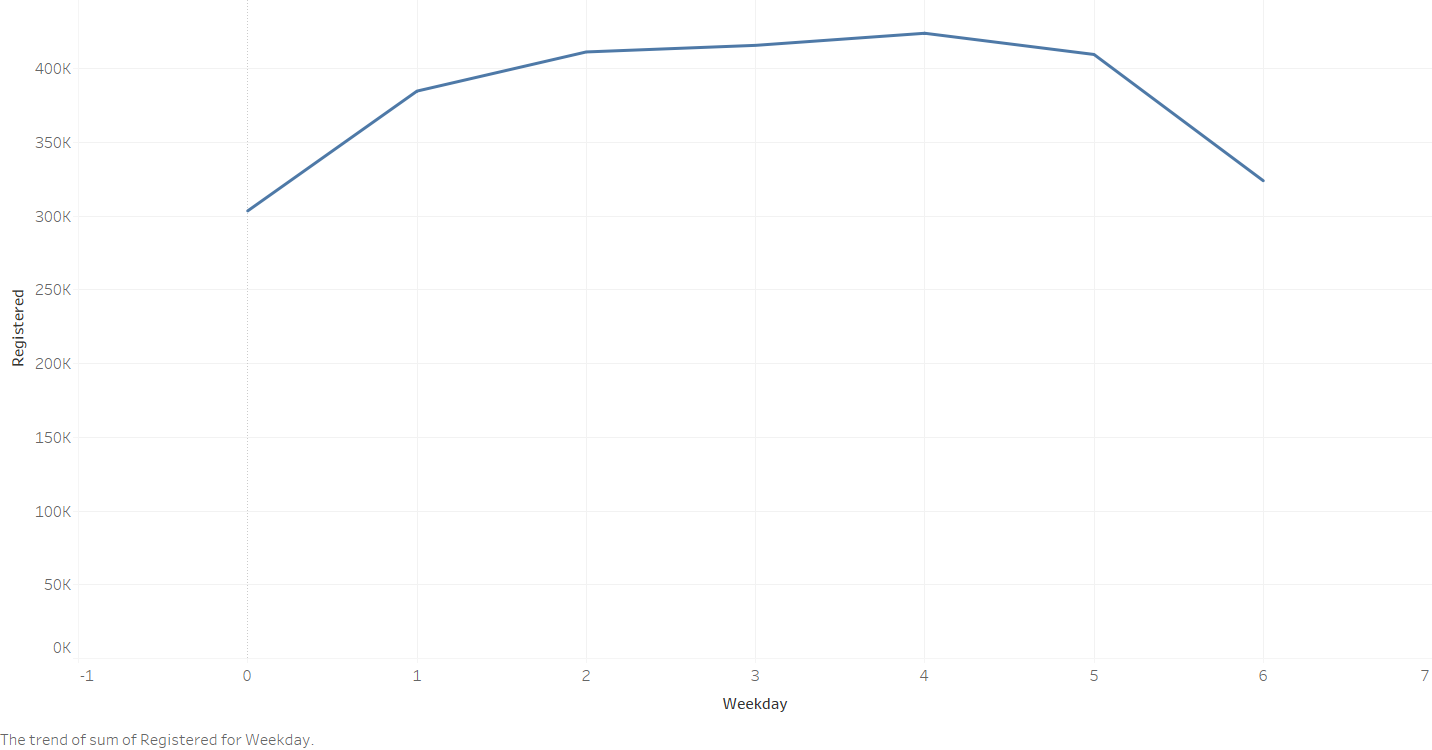
R has predicted the number of rented bike was 522.

Python has predicted the number of rented bike was 5440.

The difference of predicted bike rent is quite huge in R and Python. But as the error rate of R is quite negligible so we can take the value predicted by R.

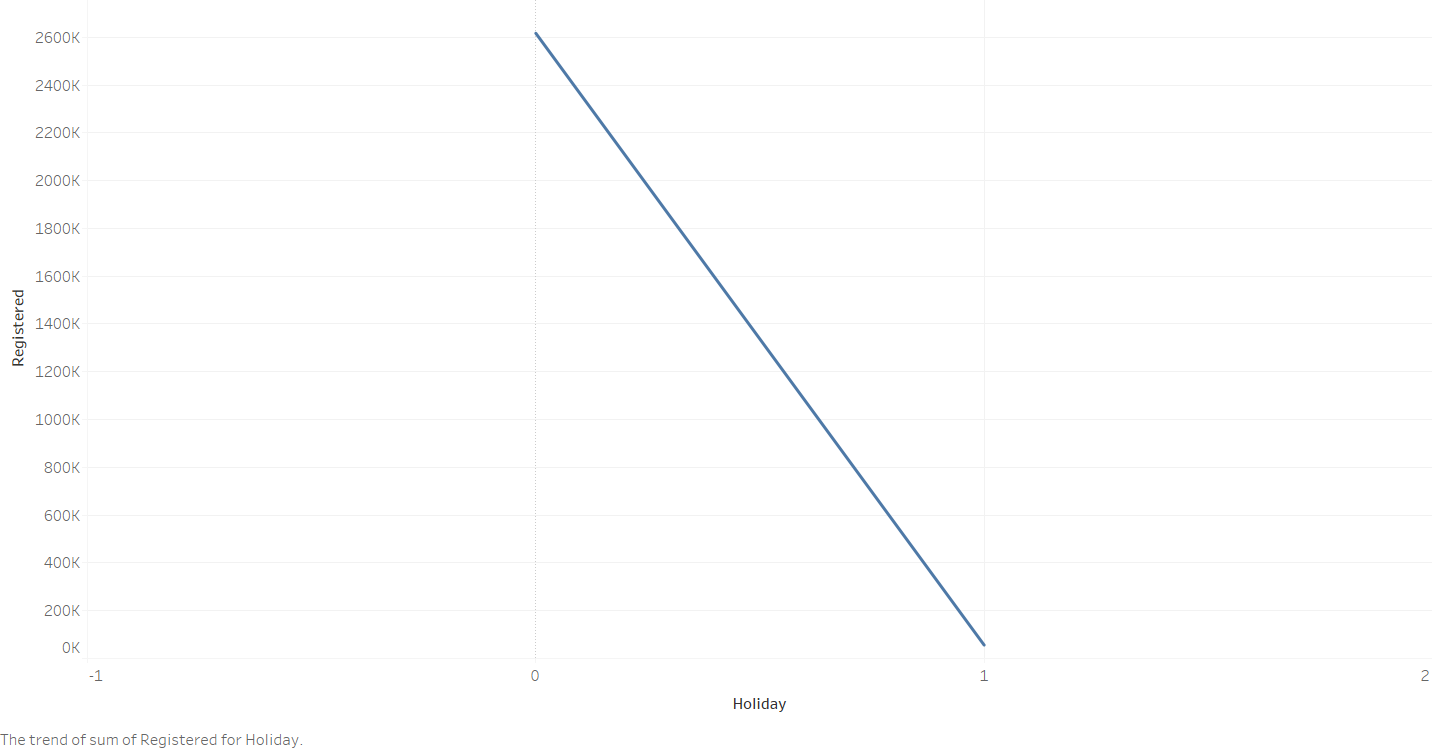
# 7 Visualization

## 7.1 count of rent weekday wise



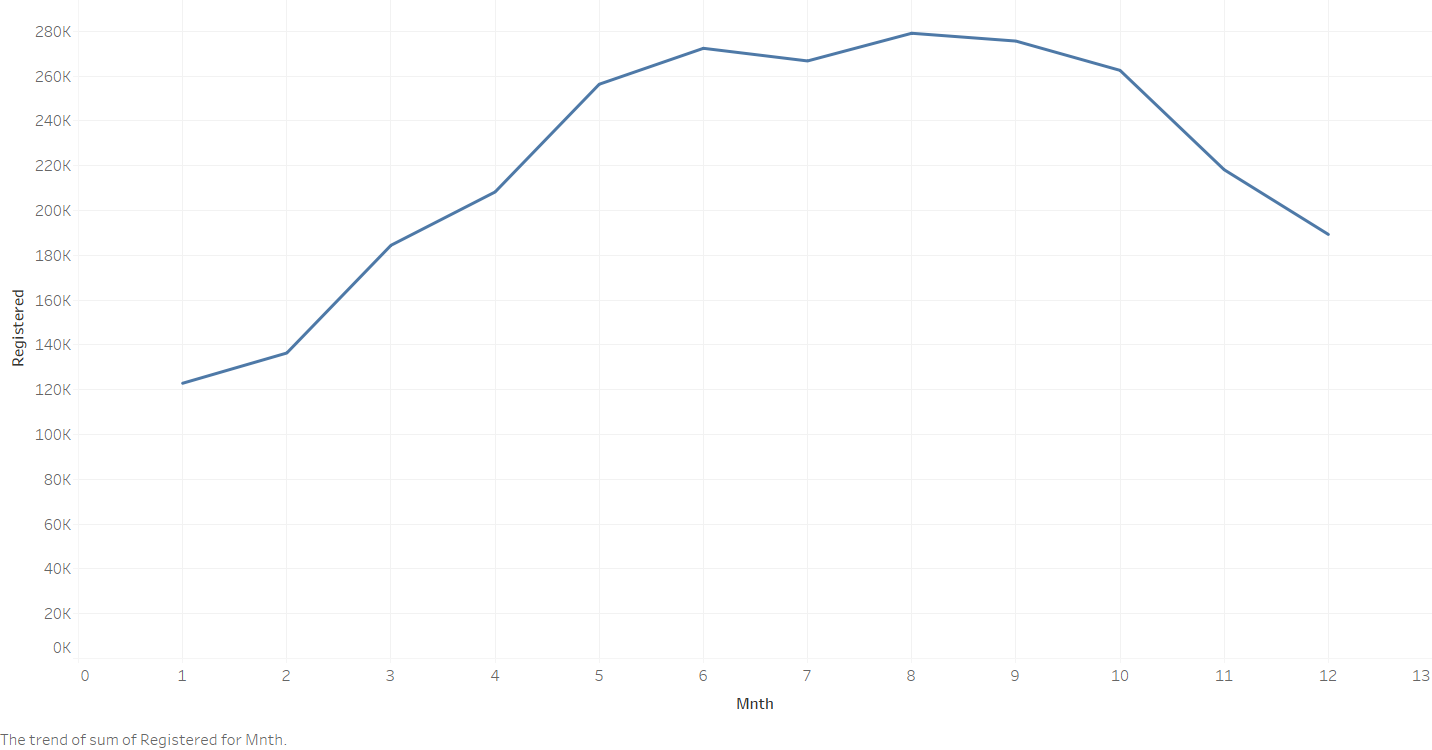
We can clearly see that the average fare amount 2015 was much higher than prior years.

## 7.2 count of rents holiday wise



Here we can clearly see that the rate rides on Fridays and it goes down from Sunday to Wednesday.

## 7.3 count of rents month wise



As the avg. fare has increased drastically in 2015 the no. of rides has reduced drastically. We have to control our fare to maintain avg. no. of rides.